Z-AXIS: From Deterministic Ground to Agentic Depth for Enterprise Evaluation

Zifan Song[†] **Mianzhi Chang** **Ziyang Liao** **Haiyan Xu**
Yutong Liu **Cairong Zhao**[✉]
School of Computer Science and Technology
Tongji University
Shanghai 201804, China
sugger@tongji.edu.cn
zhaocairong@tongji.edu.cn

Abstract

Agentic systems powered by large language models have advanced rapidly yet current architectures treat information acquisition as a homogeneous process, issuing broad searches, accumulating heterogeneous evidence, and deferring verification to post-hoc reconciliation. This neglect of acquisition-mode structure leaves factual grounding and epistemic reliability as emergent rather than engineered properties. In this paper, we formalize the ***D**eterministic-**A**gentic (**D/A**) Separation Principle*, which partitions the task-relevant information space into programmable ($\mathcal{D}$) and emergent ($\mathcal{A}$) zones, establishing $\mathcal{D}$ as both epistemic anchor and investigative compass for $\mathcal{A}$-directed exploration. We instantiate this principle as **Z-AXIS**, a three-phase cascade that front-loads deterministic grounding, deploys hierarchical wave agents for macro-to-micro deep research, and synthesizes evidence through tension-aware cross-dimensional analysis. Key mechanisms include authority-calibrated evidence fusion, D-conditioned coverage tracking, and temporal awareness for longitudinal evaluation. Extensive evaluation across multiple industries demonstrates that Z-AXIS achieves superior factual accuracy and epistemic reliability compared to existing deep research systems, with architecture-determined reliability remaining stable across a wide range of backbone scales, confirming that structured governance, not model scale, governs epistemic reliability.

1 Introduction

Agentic systems built upon large language models (LLMs) have traversed from tool-augmented generation [43, 11, 31, 14] to knowledge-intensive tasks (*e.g.*, deep research [7, 59, 23] and scientific discovery). The frontier has now shifted from capability demonstration to *guaranteed reliability*: from systems that produce impressive outputs under favorable conditions to architectures that ensure factual grounding, epistemic discipline, and information completeness as structural properties rather than hoped-for emergent behaviors. This reliability gap is especially acute in domains that interleave programmable data retrieval with open-ended investigative reasoning.

Enterprise evaluation crystallizes this challenge. This real-world decision-support task demands collecting, verifying, and synthesizing multi-source evidence (*e.g.*, industry position, competitive landscape, team capability, and growth potential) to support investment due diligence or credit assessment [53, 10]. Domain experts navigate a hierarchical cognitive process from macro-level context to micro-level detail, where each layer constrains the hypothesis space for the next. Yet current agents treat the information flow as a transparent pipe: they issue broad searches, accumulate

† Project Lead. ✉ Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

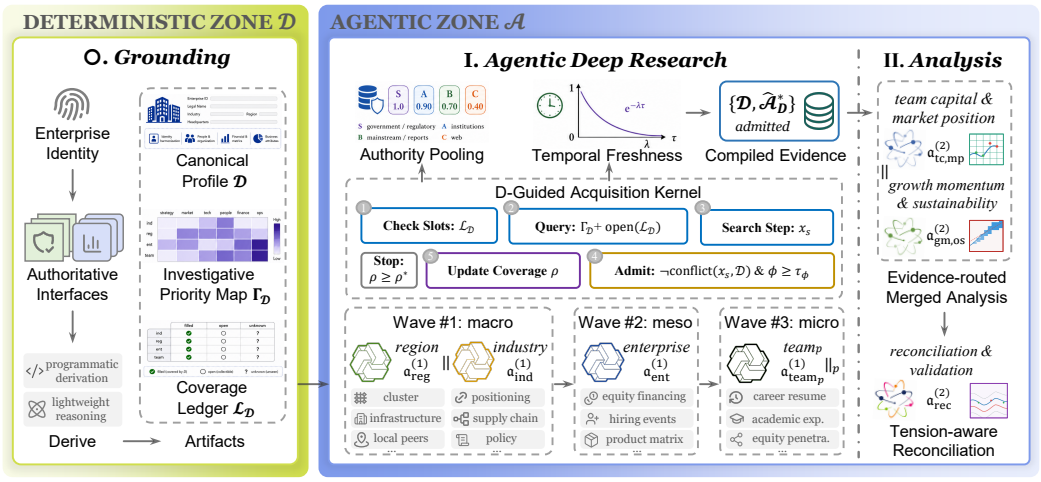

Figure 1: **Architecture of the proposed Z-AXIS system comprising three information-cascaded phases**: *deterministic grounding*, *agentic deep research*, and *structured analysis*.

heterogeneous evidence, and rely on the model to reconcile conflicts at the end, rarely distinguishing what can be deterministically obtained from what genuinely requires exploratory reasoning.

"To know what you know and what you do not know, that is true knowledge." – The Analects

We formalize this ancient intuition as the ***D**eterministic-**A**gentic (**D/A**) Separation Principle*: for deep research tasks with clear objective boundaries, the task-relevant information space $\mathcal{I}$ admits an exploitable partition $\mathcal{I} = \mathcal{D} \cup \mathcal{A}$, where $\mathcal{D} \cap \mathcal{A} = \emptyset$. The distinction lies not in whether queries carry parameters (both zones do), but in whether the *acquisition procedure* is programmable or emergent. The deterministic zone $\boldsymbol{\mathcal{D}}$ contains information whose acquisition can be specified in advance through parameterized queries, rule-based derivations, and conditional branching without agentic research involvement. The agentic zone $\boldsymbol{\mathcal{A}}$ contains information whose acquisition path emerges through reasoning over intermediate discoveries, requiring agents to decide what to investigate next, which evidence to pursue, and when to stop. This separation suggests a simple but powerful strategy: saturate $\mathcal{D}$ first at low marginal cost, then allocate agentic reasoning exclusively to $\mathcal{A}$ (high marginal value). Crucially, $\mathcal{D}$ serves a dual role: (1) *epistemic anchor*, establishing a factual horizon against which downstream agentic claims can be checked, and (2) *investigative compass*, indicating where agentic exploration should concentrate. A fact such as "core inventors are two or fewer" is not merely a database field; it simultaneously flags key-person dependency as a high-priority research direction. Enterprise evaluation is particularly favorable for this principle because structured domain databases provide abundant programmable hooks, yielding a rich $\mathcal{D}$ with industry-specific extensibility.

Sequentially, we instantiate this principle as **Z-AXIS**[1], an agentic system that implements a three-phase information cascade for enterprise evaluation, as illustrated in Figure 1. *Phase 0 — deterministic grounding* constructs a structured baseline profile through programmable queries with conditional branching, simultaneously serving as factual ground, research directive source, and completeness checklist. *Phase 1 — agentic deep research* deploys hierarchical agents at macro (region and industry landscape), meso (enterprise positioning), and micro (team-level investigation) levels that operate exclusively on $\mathcal{A}$, guided by $\mathcal{D}$-derived directives and constrained by completeness tracking. *Phase 2 — structured analysis* synthesizes research artifacts into evaluation through merged dimension analysis, cross-dimensional consistency validation, and dual-channel output generation. Each phase hands off structured artifacts to the next, ensuring that research and analysis remain decoupled and that raw search traces do not pollute downstream reasoning.

Our proposed architecture redefines the role of agents in deep research. Instead of asking agents to search everything, Z-AXIS asks agents to search only what cannot be programmatically known. Instead of relying on agents to remember every detail, it provides a deterministic baseline as an external memory and factual constraint. Instead of asking a single model to simultaneously discover,

[1]The name reflects a dual-layer design: in the research dimension, the XY plane captures $\mathcal{D}$ (ground coordinates) while the Z-axis represents $\mathcal{A}$ (emergent insights from deep exploration); in the analysis dimension, X = Region, Y = Industry, Z = Enterprise, forming a *Trinity* that grounds evaluation in complete spatial context.

verify, and evaluate, it separates acquisition modes, research roles, and analysis responsibilities across the cascade. The result is an agentic system designed not merely for impressive generation, but for evidence-grounded, completeness-aware, and structurally reliable enterprise evaluation. The core contributions of our work are threefold:

- **Principle.** We formalize the *D/A Separation Principle*, partitioning the information space into deterministic (programmable acquisition) and agentic (emergent acquisition) zones, and establish that $\mathcal{D}$ simultaneously serves as epistemic anchor and investigative compass for $\mathcal{A}$. To our knowledge, no prior work formalizes this programmable-versus-emergent information partition.
- **System.** We present Z-AXIS, an agentic system with a three-phase cascade (deterministic grounding, agentic deep research, structured analysis) and a region-industry-enterprise trinity. Key mechanisms include authority-calibrated evidence fusion, prior-driven exploration with completeness tracking, generator-evaluator consistency validation, dual-phase cognitive decoupling, and temporal awareness for longitudinal evaluation.
- **Evaluation.** We evaluate Z-AXIS on 300 enterprises across five industries against six baseline systems through seven experiments, demonstrating architecture-determined factual reliability across a 178$\times$ LLM parameter range, with an FP standard deviation of approximately 1.0 percentage point ($p = 0.321$), and the lowest composite hallucination score among all systems.

2 Deterministic-Agentic Separation Principle

Deep research is not a homogeneous information-seeking process. Given a research object under a fixed evaluation rubric, some task-relevant evidence can be acquired through programmable procedures with bounded cost and verifiable outputs, whereas other evidence requires adaptive exploration whose query path depends on intermediate findings. We formalize this observation as the *Deterministic-Agentic (D/A) Separation Principle*: reliable deep research should separate acquisition modes before reasoning begins. Deterministic acquisition establishes a verified factual ground, while agentic acquisition is reserved for residual, open-ended, and context-dependent inquiry.

Acquisition-mode separation. Let $\mathcal{I}$ denote the task-relevant information space for the target object under the evaluation rubric. We do not assume that an information item is intrinsically deterministic or agentic. Instead, the boundary depends on available interfaces, acquisition cost, and reliability requirement. A deterministic policy $\pi \in \Pi_{\mathrm{D}}$ follows pre-specified program logic, such as structured database queries, rule-based extraction, schema matching, or API calls. An agentic policy $\pi \in \Pi_{\mathrm{A}}$ chooses its next query or source based on intermediate discoveries and language-model reasoning over the current research state. For a reliability tolerance ϵ, the D/A boundary is defined as:

$$\mathcal{D} = \{x \in \mathcal{I} : \exists\, \pi \in \Pi_{\mathrm{D}} \text{ s.t. } r(\pi, x) \geq 1 - \epsilon\}, \qquad \mathcal{A} = \mathcal{I} \setminus \mathcal{D}. \tag{1}$$

where $r(\pi, x) \in [0, 1]$ denotes the reliability with which policy π acquires item x, measured as the probability of returning a validated result within admissible cost. In the sequel, $\mathcal{D}$ denotes the acquired deterministic ground under this reliability requirement, and $\mathcal{A}$ denotes the residual information space requiring agentic acquisition. Essentially, the D/A boundary is target-conditioned, resource-bounded, and interface-dependent. For example, a firm's registered capital, legal representative, patent statistics, financing history, and bidding records fall into $\mathcal{D}$ when structured sources are available, but may shift to $\mathcal{A}$ when it can only be inferred from noisy web pages or unstructured reports. Conversely, founder reputation, affiliated-entity risks, technological defensibility, and market-positioning narratives usually remain in $\mathcal{A}$ because their relevant queries depend on what is discovered during investigation.

Why separation helps. The deterministic baseline does not merely provide background information. It reshapes the downstream research problem by serving two coupled roles: an *investigative compass* that prioritizes where agentic exploration should be spent, and an *epistemic anchor* that determines which exploratory findings can be trusted. After deterministic grounding, agentic research should not search all of $\mathcal{A}$ uniformly. Instead, the deterministic ground induces a priority function $\gamma_{\mathcal{D}} : \mathcal{A} \to \mathbb{R}_{\geq 0}$, measuring the investigative priority of an A-zone item x under $\mathcal{D}$. This priority score subsumes both unresolved information needs and deterministic signals that trigger deeper investigation. The D-conditioned agentic target space is therefore

$$\mathcal{A}_{\mathcal{D}} = \{x \in \mathcal{A} : \gamma_{\mathcal{D}}(x) \geq \tau_{\gamma}\}. \tag{2}$$

Here, τ_γ is the priority threshold for agentic exploration. This formulation captures a key property of D/A separation: $\mathcal{D}$ does not merely shrink the search space from $\mathcal{I}$ to $\mathcal{A}$; it reorganizes $\mathcal{A}$ into a directed research agenda. A deterministic observation such as "many tenders won, but almost all in the same prefecture-level city" is not merely evidence of sales traction; it raises the priority of investigating regional policy dependence, relationship-driven procurement, and cross-region replicability. Likewise, when hiring expands rapidly while financing or order records do not move in tandem, the system prioritizes investigation into hidden financing or delivery overload. Thus, deterministic evidence converts static facts into directional signals for agentic inquiry.

The same deterministic ground also constrains admissibility. Let $\phi(x)$ denote the source-confidence score (specified in Section 3.2) of an A-zone finding x. The admissible agentic evidence is defined as:

$$\mathcal{A}^\star_{\mathcal{D}} = \{x \in \mathcal{A}_{\mathcal{D}} : \neg\ \mathrm{conflict}(x, \mathcal{D})\ \wedge\ \phi(x) \geq \tau_\phi\}\,. \tag{3}$$

Only findings in $\mathcal{A}^\star_{\mathcal{D}}$ are allowed to enter downstream evaluation. The consistency constraint $\neg\ \mathrm{conflict}(x, \mathcal{D})$ ensures that exploratory claims do not contradict the deterministic ground, while the confidence threshold τ_ϕ prevents weakly sourced claims from being admitted merely because they are plausible. Therefore, D/A separation helps not simply because deterministic acquisition is cheaper, but because deterministic grounding transforms open-ended exploration over $\mathcal{I}$ into prioritized search over $\mathcal{A}_{\mathcal{D}}$, followed by admissibility filtering against $\mathcal{D}$. In this sense, $\mathcal{D}$ serves simultaneously as the factual horizon and the investigative compass for $\mathcal{A}$.

Budgeted D-first advantage. The budget argument is secondary to the structural effect above, but it explains why D/A separation is especially attractive in practical research settings. Let B denote the total research budget and C_{D} the cost of constructing $\mathcal{D}$. D-first acquisition is favorable when the direct value of deterministic grounding, together with the value of agentic exploration over the narrowed target space $\mathcal{A}_{\mathcal{D}}$ using the remaining budget $B - C_{\mathrm{D}}$, exceeds the value of spending the full budget B on unguided exploration over $\mathcal{A}$. This regime is likely when deterministic acquisition is cheap, reliable, and directive-rich. Enterprise evaluation is precisely such a setting: registries, patent databases, financing records, qualification databases, recruitment platforms, and bidding databases provide abundant programmable hooks, yielding a high-value deterministic ground at low marginal cost while directing agentic effort toward the residual questions that matter most.

3 Z-AXIS System: Three-Phase Cascade

We instantiate the D/A Separation Principle as **Z-AXIS**, an agentic system for enterprise evaluation. Following Eq. 1, we first acquire the deterministic ground $\mathcal{D}$ through programmable invocation, and subsequently allocate agentic exploration to the residual space $\mathcal{A}$. To achieve this, Z-AXIS adopts a three phase information cascade. Phase 0 constructs a deterministic grounding artifact. Phase 1 deploys role specialized research agents: *industry* $\mathfrak{a}^{(1)}_{\mathrm{ind}}$, *region* $\mathfrak{a}^{(1)}_{\mathrm{reg}}$, *enterprise* $\mathfrak{a}^{(1)}_{\mathrm{ent}}$, and *team*$_p$ $\mathfrak{a}^{(1)}_{\mathrm{team},p}$ for key person p. Phase 2 deploys analysis agents $\mathfrak{a}^{(2)}_{\mathrm{tc,mp}}$, $\mathfrak{a}^{(2)}_{\mathrm{gm,os}}$, and $\mathfrak{a}^{(2)}_{\mathrm{rec}}$ to synthesize evidence and resolve cross dimensional conclusions. We denote the final enterprise evaluation by $\mathcal{E}$, use $G(\mathcal{E})$ to describe the evaluation relevant information accounted for by the cascade, and use $I(\cdot\,;\cdot)$ for mutual information. This gives the following architectural decomposition:

$$G(\mathcal{E}) = \underbrace{I(\mathcal{E};\,\mathcal{D})}_{\text{Phase 0: grounding}} + \underbrace{I(\mathcal{E};\,\mathfrak{a}^{(1)} \mid \mathcal{D})}_{\text{Phase 1: exploration}} + \underbrace{I(\mathcal{E};\,\mathfrak{a}^{(2)} \mid \mathcal{D}, \mathfrak{a}^{(1)})}_{\text{Phase 2: analysis}}. \tag{4}$$

Crucially, each phase emits structured artifacts, so research and analysis remain separated in context and raw search traces do not directly enter final reasoning.

3.1 Phase 0: Deterministic Grounding

Phase 0 aims to construct a deterministic grounding artifact $(\mathcal{D}, \{\Gamma_{\mathcal{D},r}, \mathcal{L}_{\mathcal{D},r}\}_r)$. Here, $\mathcal{D}$ is the deterministic ground, $\Gamma_{\mathcal{D},r} = \{(x, \gamma_{\mathcal{D},r}(x)) : x \in \mathcal{A}, \gamma_{\mathcal{D},r}(x) \geq \tau_\gamma\}$ is the D-induced directive set for Phase 1 role r, and $\mathcal{L}_{\mathcal{D},r}$ is the D-prefilled coverage ledger for that role. The priority threshold τ_γ is shared across roles, while $\gamma_{\mathcal{D},r}$ captures role specific investigative priority as motivated by Eq. 2.

Specifically, Phase 0 turns an enterprise target into a compact factual substrate. We first resolve the canonical enterprise identity and harmonize registry level records into a unified profile. Anchored by this profile, our system retrieves structured evidence from authoritative domain interfaces covering

corporate status, intellectual property, financing, qualifications, procurement, recruitment, standards participation, and talent signals. Subsequently, programmatic derivations and lightweight model reasoning convert these fields into evaluation relevant signals, including technology trajectory, financing stage, procurement footprint, and key person dependency. Consequently, deterministic grounding enters the cascade as factual substrate, research guidance, and coverage baseline.

3.2 Phase 1: Agentic Deep Research

Phase 1 performs agentic acquisition over the D-conditioned target space $\mathcal{A}_{\mathcal{D}}$ defined in Eq. 2. We organize research as a macro→meso→micro hierarchy. Wave 1 runs industry and region agents in parallel. Wave 2 runs the enterprise agent conditioned on Wave 1. Wave 3 runs team agents in parallel, one for each key person. This design follows the cognitive structure of enterprise evaluation: industry and regional context clarify the opportunity space, enterprise investigation locates the target within that space, and team investigation tests whether the organization can sustain the inferred position. Significantly, each agent receives only the context required by its role, rather than all upstream traces.

Scoped cascade and admissible evidence. Algorithm 1 shows the per-agent exploration mechanism. Formally, the agentic topology is:

$$\begin{aligned}
\mathcal{R}_{\mathrm{W1}} &= \mathfrak{a}^{(1)}_{\mathrm{ind}}(\mathcal{D}, \Gamma_{\mathcal{D},\mathrm{ind}}, \mathcal{L}_{\mathcal{D},\mathrm{ind}}) \;\|\; \mathfrak{a}^{(1)}_{\mathrm{reg}}(\mathcal{D}, \Gamma_{\mathcal{D},\mathrm{reg}}, \mathcal{L}_{\mathcal{D},\mathrm{reg}}) \,, \\
\mathcal{R}_{\mathrm{W2}} &= \mathfrak{a}^{(1)}_{\mathrm{ent}}(\mathcal{D}, \mathcal{R}_{\mathrm{W1}}, \Gamma_{\mathcal{D},\mathrm{ent}}, \mathcal{L}_{\mathcal{D},\mathrm{ent}}) \,, \\
\mathcal{R}_{\mathrm{W3}} &= \left\{ \mathfrak{a}^{(1)}_{\mathrm{team},p}(\mathcal{D}, \mathcal{R}_{\mathrm{W2}}, \Gamma_{\mathcal{D},\mathrm{team},p}, \mathcal{L}_{\mathcal{D},\mathrm{team},p}) \right\}_{p\in\mathcal{P}} .
\end{aligned} \tag{5}$$

This scoped design prevents raw macro narratives from entering team investigation. For each wave $m \in \{1,2,3\}$, we admit only evidence satisfying $\widehat{\mathcal{A}}^{\star}_{\mathcal{D},m} = \{x \in \mathcal{R}_{\mathrm{W}m} : \neg\ \mathrm{conflict}(x,\mathcal{D}) \wedge \phi(x) \geq \tau_\phi\}$, and aggregate the admitted evidence as $\widehat{\mathcal{A}}^{\star}_{\mathcal{D}} = \cup_{m=1}^{3} \widehat{\mathcal{A}}^{\star}_{\mathcal{D},m}$, consistent with Eq. 3.

Coverage, hallucination attenuation, and evidence fusion. For each role r, $\mathcal{L}_{\mathcal{D},r}$ records which evidence slots are already filled by $\mathcal{D}$ and which remain open for A-zone exploration. The ledger is semi-fixed: a frozen, human-authored slot ontology defines a common spine, while deterministic rules over $\mathcal{D}$ activate an enterprise-specific fringe. Agents may formulate queries and fill evidence slots, but cannot modify the schema; coverage is computed over the instantiated ledger. At step s, the local coverage ratio is $\rho_{r,s} = |\,\mathrm{fill}(\mathcal{L}_{\mathcal{D},r}; \mathcal{D}\cup\mathcal{R}^{(s)})|/|\mathcal{L}_{\mathcal{D},r}|$, where $\mathcal{R}^{(s)}$ is the loop local report state. Every k steps, each agent prunes redundancy, audits consistency against $\mathcal{D}$, and re ranks remaining gaps by marginal value. Exploration stops when $\rho_{r,s} \geq \rho^*$ or the maximum step count $S_{\max}$ is reached. Sequentially, we employ a progressive discipline to attenuate hallucination. First, deterministic facts define an inviolable factual horizon, so any A-zone finding that contradicts $\mathcal{D}$ is rejected. Second, every non deterministic claim must include source attribution and timestamp information. Third, unavailable dimensions are marked as unknown rather than completed by speculation. Finally, Phase 2 checks whether dimension level conclusions remain logically compatible before final delivery. To support this admissibility test, we compute source reliability by combining authority and temporal freshness:

$$\phi = \underbrace{\left(1 - \prod (1 - w)\right)}_{\text{authority pooling}} \cdot \underbrace{e^{-\lambda \cdot \tau}}_{\text{temporal decay}} \quad , \tag{6}$$

where the product ranges over corroborating sources, $w \in [0,1]$ is the reliability weight assigned by authority tier (S: government and regulatory databases, w=1.0; A: authoritative industry institutions, w=0.90; B: mainstream media and research reports, w=0.70; C: general web sources, w=0.40), τ is elapsed time since publication, and λ is a domain adapted decay rate. For instance, one A + one B verification yields $\phi \approx 1 - (1-0.90)(1-0.70) = 0.97$ before temporal decay. Only findings satisfying $\phi(x) \geq \tau_\phi$ and D consistency enter $\widehat{\mathcal{A}}^{\star}_{\mathcal{D}}$.

3.3 Phase 2: Structured Analysis

Phase 2 converts $\mathcal{D}$ and admitted A-zone evidence into the final enterprise evaluation $\mathcal{E}$. We evaluate each enterprise along four dimensions. Specifically, (1) *team capital* measures leadership depth, team stability, complementarity, key person dependency, talent pipeline, and organizational learning capacity; (2) *market position* measures ecological niche, competitive differentiation, customer concentration, supply chain position, and alignment with industry trajectory; (3) *growth momentum* measures scaling potential through revenue signals, patent trajectory, product pipeline, market

Algorithm 1 Phase 1 D-Guided Acquisition Kernel

Require: agent $\mathfrak{a}_r^{(1)}$, scoped context, $\mathcal{D}$, $\Gamma_{\mathcal{D},r}$, $\mathcal{L}_{\mathcal{D},r}$
Ensure: source-attributed report $\mathcal{R}^{(s)}$
$\mathcal{R}^{(0)} \leftarrow \emptyset$
$\rho_{r,0} \leftarrow |\,\mathrm{fill}(\mathcal{L}_{\mathcal{D},r}; \mathcal{D})|/|\mathcal{L}_{\mathcal{D},r}|$
for $s = 1, \ldots, S_{\max}$ **while** $\rho_{r,s-1} < \rho^*$ **do**
  derive queries from $\Gamma_{\mathcal{D},r}$ and $\mathcal{L}_{\mathcal{D},r} \setminus \mathrm{fill}(\mathcal{L}_{\mathcal{D},r}; \mathcal{D} \cup \mathcal{R}^{(s-1)})$
  search A-zone claims $\mathcal{X}_s$ with source attribution and timestamps
  $\mathcal{R}^{(s)} \leftarrow \mathcal{R}^{(s-1)} \cup \{x \in \mathcal{X}_s : \neg\,\mathrm{conflict}(x, \mathcal{D}) \wedge \phi(x) \geq \tau_\phi\}$
  update $\mathcal{L}_{\mathcal{D},r}$ using $\mathcal{D} \cup \mathcal{R}^{(s)}$; keep unfilled slots unknown
  if $s \bmod k = 0$: prune redundancy; re-rank open slots by $\gamma_{\mathcal{D},r}$
  $\rho_{r,s} \leftarrow |\,\mathrm{fill}(\mathcal{L}_{\mathcal{D},r}; \mathcal{D} \cup \mathcal{R}^{(s)})|/|\mathcal{L}_{\mathcal{D},r}|$
return $\mathcal{R}^{(s)}$

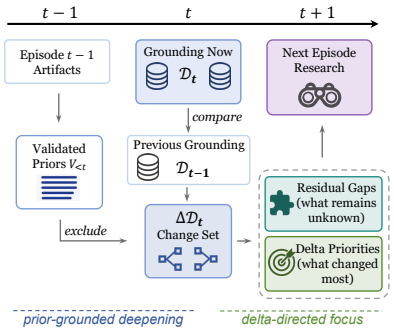

Figure 2: **Key mechanism of temporal awareness.**

expansion, hiring, and financing momentum; (4) *operational sustainability* measures structural resilience through financial health, compliance, litigation exposure, governance quality, and business model durability. Crucially, Phase 2 agents do not acquire new evidence. They synthesize admitted evidence under explicit dimensional boundaries.

Evidence routed merged analysis. To promote coherent reasoning, we merge related dimensions into two inference passes while routing only the evidence scope required by each role. Team capital and market position are analyzed together because person-level capability often explains competitive positioning. Growth momentum and operational sustainability are analyzed together because both concern longitudinal viability and structural resilience. The resulting pipeline is:

$$\widehat{\mathcal{E}}_{\mathrm{tc,mp}} = \mathfrak{a}^{(2)}_{\mathrm{tc,mp}}\left(\mathcal{D}, \widehat{\mathcal{A}}^{\star}_{\mathcal{D},1}, \widehat{\mathcal{A}}^{\star}_{\mathcal{D},2}, \widehat{\mathcal{A}}^{\star}_{\mathcal{D},3}\right), \quad \widehat{\mathcal{E}}_{\mathrm{gm,os}} = \mathfrak{a}^{(2)}_{\mathrm{gm,os}}\left(\mathcal{D}, \widehat{\mathcal{A}}^{\star}_{\mathcal{D},1}, \widehat{\mathcal{A}}^{\star}_{\mathcal{D},2}\right)$$
$$\mathcal{E} = \mathfrak{a}^{(2)}_{\mathrm{rec}}\left(\widehat{\mathcal{E}}_{\mathrm{tc,mp}}, \widehat{\mathcal{E}}_{\mathrm{gm,os}}; \mathcal{D}, \widehat{\mathcal{A}}^{\star}_{\mathcal{D}}\right). \tag{7}$$

The first pass receives all three evidence waves, while the second pass deliberately excludes $\widehat{\mathcal{A}}^{\star}_{\mathcal{D},3}$ unless team information has already been reflected in $\mathcal{D}$ or Wave 2 findings. Consequently, person narratives do not contaminate trajectory and sustainability judgments.

Tension aware reconciliation. The reconciliation agent $\mathfrak{a}^{(2)}_{\mathrm{rec}}$ audits the two partial evaluations against each other and against the admitted evidence base. It does not average dimension scores. Instead, it identifies tensions, distinguishes evidential conflicts from rhetorical inconsistency, and resolves cross dimensional conclusions through explicit evidence chains. For example, a strong market position rating paired with weak team capital must be explained as externally supported market access, founder independent organizational capability, or an unresolved fragility. The final output $\mathcal{E}$ therefore contains ratings, rationales, supporting evidence, uncertainty notes, and contradiction resolutions.

3.4 Temporal Awareness

Enterprise evaluation is longitudinal. At episode t, Z-AXIS reuses the same cascade with current deterministic ground $\mathcal{D}_t$ and admitted evidence $\widehat{\mathcal{A}}^{\star}_{\mathcal{D}_t,t}$. Instead of naive replay, temporal evaluation exploits two complementary mechanisms as depicted in Figure 2. *Prior-grounded deepening*: validated findings from earlier episodes raise the research baseline, so that agentic exploration at t probes residual questions at greater depth rather than retracing covered ground. *Delta-directed focus*: changes in the deterministic ground $\Delta\mathcal{D}_t = \mathcal{D}_t \ominus \mathcal{D}_{t-1}$ generate fresh investigation priorities, directing research toward emergent risks and opportunities. Formally, the verified prior set and the episode-t target space are:

$$\mathcal{V}_{<t} = \left\{x \in \widehat{\mathcal{A}}^{\star}_{\mathcal{D}_{t-1}} : \neg\,\mathrm{conflict}(x, \mathcal{D}_t) \wedge \phi_t(x) \geq \tau_\phi\right\},$$
$$\mathcal{A}^{(t)}_{\mathcal{D}_t} = \underbrace{(\mathcal{A}_{\mathcal{D}_t} \setminus \mathcal{V}_{<t})}_{\text{residual gaps}} \cup \underbrace{\{x \in \mathcal{A} : \gamma_{\Delta\mathcal{D}_t}(x) \geq \tau_\gamma\}}_{\text{delta-induced priorities}}. \tag{8}$$

The residual-gap term narrows the research frontier by excluding dimensions already established by verified priors, while the delta-induced term promotes newly relevant directions triggered by

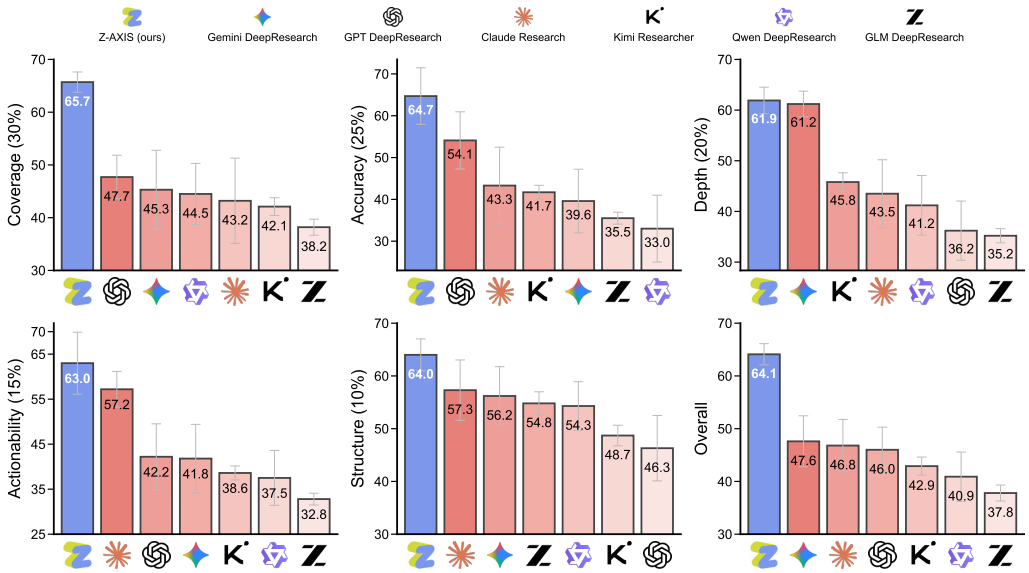

Figure 3: **Dual-channel system comparison** across seven systems and five dimensions.

Table 1: **Cross-industry dual-channel assessment I**: TMT/AI; Semiconductor. Overall = $0.30\times$cov$+$ $0.25\times$acc $+$ $0.20\times$dep $+$ $0.15\times$act $+$ $0.10\times$str. **Bold**/underline = best/second.

System	*TMT/AI*						*Semiconductor*					
	cov.	**acc.**	**dep.**	**act.**	**str.**	**ov.**	**cov.**	**acc.**	**dep.**	**act.**	**str.**	**ov.**
Gemini DR	42.2	34.1	56.4	44.7	52.3	44.4	61.3	46.2	**64.1**	55.3	59.4	57.0
GPT DR	38.6	45.2	25.7	32.4	30.9	36.0	43.2	41.4	37.1	40.5	53.3	42.1
Claude Res.	42.1	29.3	46.6	52.4	**64.8**	43.6	42.4	48.3	53.0	56.2	57.4	49.6
Kimi Res.	45.3	44.8	50.2	42.3	53.6	46.5	39.4	39.1	42.6	35.3	46.2	40.0
Qwen DR	35.4	21.7	28.3	42.1	46.6	32.7	58.1	47.2	43.2	53.4	53.1	51.2
GLM DR	38.6	35.4	35.3	31.8	55.7	37.8	42.3	40.1	41.6	37.2	58.3	42.4
Z-AXIS (ours)	**68.3**	**63.7**	**64.1**	**66.2**	62.8	**65.4**	**62.7**	**65.3**	62.4	**58.6**	**63.1**	**62.7**

factual changes (*e.g.*, patent acceleration, key-person departure, financing events, bidding shifts, or litigation emergence). As episodes accumulate, the system transitions from broad initial research to focused monitoring: deterministic cost reduces to delta computation, A-zone effort concentrates on residual gaps and emergent priorities, and the cumulative evidence base deepens longitudinal evaluation through progressively richer prior grounding.

4 Experiments

4.1 Implementation Details

Setup. Z-AXIS deploys role-specialized agents with backbone LLMs spanning 9B to 1.6T (178$\times$ range). Research agents combine multi-source search, D-conditioned completeness tracking, and sandboxed code execution; analysis agents share code execution and cross-dimensional validation without external access. Evaluation: 300 enterprises across five industries (TMT/AI, semiconductor, biopharma, advanced manufacturing, traditional industry), stratified by stage and scale, with 25-enterprise expert blind subset. Baselines: Gemini DR, GPT DR, Claude Research, Qwen DR, GLM DR, Kimi Researcher. The benchmark comprises 78 TMT/AI, 66 semiconductor, 60 biopharma, 60 advanced manufacturing, and 36 traditional industry enterprises, stratified by development stage (28% early, 45% growth, 27% mature) and headcount (39% small, 43% medium, 18% large). Seventy percent are registered in mainland China, and the evaluation uses an April 2026 snapshot. All six baselines ran under their default configurations without baseline-specific prompt tuning. All seven systems received the same canonical deterministic profile, task instruction, evaluation rubric, and output schema. The profile contained factual fields only, without analysis or scoring answers.

Table 2: **Cross-industry dual-channel assessment II**: Biopharma; Advanced Manufacturing.

System	*Biopharma*						*Advanced Manufacturing*					
	cov.	**acc.**	**dep.**	**act.**	**str.**	**ov.**	**cov.**	**acc.**	**dep.**	**act.**	**str.**	**ov.**
Gemini DR	47.3	52.1	**65.2**	48.3	**63.4**	53.8	55.3	55.1	**65.0**	45.3	50.6	55.2
GPT DR	48.2	53.4	33.1	27.6	41.3	42.7	50.3	65.2	55.1	42.4	65.3	55.3
Claude Res.	38.3	38.1	29.3	56.4	59.1	41.2	58.1	60.4	40.2	60.4	40.1	53.6
Kimi Res.	43.6	43.5	47.8	40.7	50.8	44.7	44.2	43.2	49.2	39.6	52.4	45.1
Qwen DR	43.1	30.4	47.2	31.3	56.3	40.3	35.6	10.3	45.1	35.2	55.3	33.1
GLM DR	37.8	33.8	32.4	32.1	52.6	36.3	38.2	37.3	37.9	34.5	54.4	39.0
Z-AXIS (ours)	**65.1**	**61.3**	60.6	**62.9**	61.2	**62.5**	**68.1**	**66.4**	64.3	**64.7**	**66.1**	**66.2**

Table 3: **Objective epistemic reliability**. FP/FR: fact precision/recall; H_1: attribution, H_2: grounding, H_3: calibration, H_4: contradiction (%). cH: composite. *Profile*: epistemic feature.

System	Profile	FP ↑	FR ↑	H_1 ↓	H_2 ↓	H_3 ↑	H_4 ↓	cH ↓
Gemini DR	*ungrounded breadth*	91.8	63.2	22.4	4.8	72.1	12.3	16.9
GPT DR	*unattributed precision*	94.1	61.5	16.0	2.6	79.8	8.1	11.7
Claude Res.	*conservative sourcing*	93.4	55.8	12.1	3.1	84.2	5.8	9.2
Kimi Res.	*moderate governance*	92.3	56.3	18.7	3.5	76.5	9.4	13.8
Qwen DR	*poorly calibrated*	90.2	52.4	23.1	5.6	68.3	10.7	17.8
GLM DR	*ungoverned*	88.5	47.8	26.8	6.2	64.1	14.6	20.9
Z-AXIS (ours)	*anchor-governed cascade*	**95.1**	**71.2**	**8.2**	**1.9**	**87.4**	**4.2**	**6.7**

Evaluation protocol. Dual-channel. *Subjective*: equal-weight blend of four LLM judges (Opus 4.6, GPT-5.4, GLM-5, Qwen-3.5-Plus) and five domain experts, scoring coverage (30%), accuracy (25%), depth (20%), actionability (15%), structure (10%). *Objective*: a uniform pipeline verifies claims against an independent fact base, separate from inference-time sources, yielding FP, FR, H_1 (attribution), H_2 (grounding), H_3 (calibration), and H_4 (contradiction); $\text{cH} = (H_1 + H_2 + (100 - H_3) + H_4)/4$.

4.2 Main results

Dual-channel cross-industry assessment. Tables 1 and 2 report per-industry scores; Figure 3 aggregates system-level comparison. Z-AXIS ranks first unanimously (+16.5 overall): coverage (+18.0), accuracy (+10.6), actionability (+5.9), structure (+6.7), depth (+0.7). Gaps trace to $\mathcal{D}$'s dual roles: investigative compass directs enterprise-specific exploration (coverage, accuracy), while epistemic anchor enforces admissibility filtering (depth, actionability). Gemini ranks 2nd in depth but only 39.6 accuracy—data access without enforcement fails to prevent reasoning from incorrect premises.

Objective epistemic reliability assessment. Table 3 reports fact-verification on 300 enterprises. FP is uniformly high (88.5 to 95.1%) while epistemic metrics span 3 to 5× (H_1: 8.2 to 26.8%), revealing differentiation through evidence governance, not generation quality. The FR gap (+8.0 over Gemini) is 8× the FP gap (+1.0 over GPT), confirming $\mathcal{D}$-zone pre-loading expands verified coverage rather than per-claim precision. H_2 = 1.9% is 27% below the next baseline, validating the anchor mechanism. A ranking inversion: Gemini 2nd by expert but 5th by cH, as broad retrieval (FR 63.2) masks the worst H_1 among top-five systems, *i.e.*, a coverage-without-grounding pattern invisible to human judges. The advantage is architectural.

4.3 Analytical Study

Zone reliability and mechanism ablation. Claims exhibit a 41.6 pp verifiability gradient across D, bridge, and A, motivating differentiated evidence governance. Table 4 reveals two complementary control paths: D-conditioned guidance and ledger tracking direct exploration toward unresolved evidence, while authority calibration, temporal decay, and admissibility constraints govern which findings can enter downstream analysis. The subsystems factorize in their dominant effects: weakening coverage controls primarily reduces FR, whereas relaxing reliability constraints sharply increases hallucination with little recall gain. The ledger variants further explain the value of its semi-fixed design: a fixed spine alone misses enterprise-specific signals, while freely generated slots introduce schema drift that undermines coverage consistency and auditability. Phase 2 reconciliation provides an additional safeguard by resolving cross-dimensional tensions without further evidence acquisition. In contrast, D-only grounding yields safe-but-shallow output: its low hallucination scores largely reflect abstention rather than comprehensive understanding. Together,

Table 4: **Mechanism ablation.** Each agentic variant modifies one mechanism. Audit viability denotes decision-grade suitability. $\Gamma_{\mathcal{D}}$: D-directives; $\mathcal{L}_{\mathcal{D}}$: coverage ledger; ϕ: authority fusion (Eq. 6).

Configuration	FP↑	FR↑	H_1↓	H_2↓	cH↓	Audit Viability
Z-AXIS (full)	95.1	71.2	8.2	1.9	6.7	pass
w/o $\Gamma_{\mathcal{D}}$ directives	94.1	57.3	12.6	3.4	11.6	degraded
w/o $\mathcal{L}_{\mathcal{D}}$ coverage	93.8	62.7	11.3	2.9	10.3	degraded
ledger spine-only (fringe disabled)	94.3	66.8	10.1	2.4	8.9	**pass**
ledger free-generated (LLM slots)	93.2	63.5	12.8	3.6	12.0	*fail*
w/o authority weighting in ϕ	91.6	70.4	17.3	5.8	14.8	*fail*
w/o temporal decay ($\lambda = 0$)	94.5	71.0	9.4	3.1	9.4	degraded
w/o τ_ϕ admission	93.0	70.6	12.6	3.2	10.8	degraded
w/o D-consistency gate	89.1	70.8	21.6	6.8	18.2	*fail*
w/o phase-2 reconciliation	93.4	69.8	10.3	2.6	11.3	degraded
D-only grounding (w/o phase 1)	96.8	24.3	4.9	1.2	3.8	*fail*

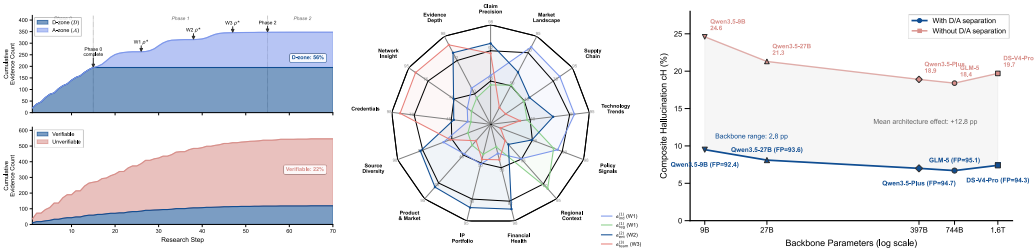

Figure 4: **Analytical study**. (a) D/A cascade (56% vs. 22% verifiable). (b) Wave agent profiles (3 quality + 10 scope axes). (c) Backbone robustness: cH ±D/A (9B to 1.6T).

these results suggest that reliable deep research requires not merely access to deterministic facts, but their active use to guide exploration, regulate evidence admission, and reconcile conclusions.

D/A cascade and wave profiles. Figure 4(a) traces D/A separation over 70 steps: Phase 0 front-loads 195 D-zone items (steps 1 to 15), then $\Gamma_{\mathcal{D}}$-directed exploration adds 154 A-zone items with wave convergence at W1 (26), W2 (38), W3 (47), yielding 56% D-zone vs. naive 22% verifiable despite 36% fewer items (2.6× advantage). (b) shows non-overlapping territorial dominance: W2 leads enterprise axes by 25 to 45 pp, W3 dominates Credentials (90.6) and Network Insight (87.8), W1 provides macro signals (Market Landscape 85.2, Technology Trends 83.6); quality axes rank orthogonally (W2 highest Claim Precision 93.7, W3 highest Evidence Depth 92.3), validating non-substitutability.

Architecture-to-backbone robustness. Figure 4(c) shows cH across five backbones (9B to 1.6T). With D/A, cH varies only 2.8 pp (FP 2.7 pp, $p = 0.321$); without it, cH ranges from 18.4 to 24.6 (mean +12.8 pp). A 27B backbone with D/A (cH = 8.1) outperforms 1.6T without it (cH = 19.7): structured governance with a modest backbone surpasses a far stronger model lacking architectural discipline.

Temporal research dynamics. Figure 5 and Table 5 illustrate how temporal awareness shifts exploration from broad acquisition to selective evidence updates (Eq. 8). At t_1, uneven authoritative coverage exposes substantial evidence gaps, particularly in Financial. Rather than repeating the initial search, prior-grounded t_2 improves evidence quality across dimensions by acquiring stronger sources, pruning weak evidence, and directing exploration toward unresolved gaps. At t_3, event-driven changes further reshape research priorities, concentrating nearly 90% of new evidence in Talent, Financial, and IP while limiting unnecessary revisits elsewhere. This progression highlights two

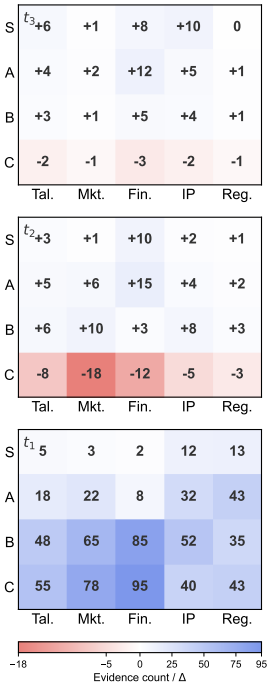

Figure 5: **Temporal dynamics**. t_1: baseline. t_2: prior-grounded. t_3: delta-directed (90% in 3 dims).

Table 5: **Temporal evidence evolution.** Evidence counts by authority tier (S+A/B/C). S+A: authoritative; B: intermediate; C: prunable evidence. t_1 gives absolute counts; t_2 and t_3 give changes from the preceding episode. Bold/underlined S+A values denote the largest/second-largest per episode.

Dimension	t_1 **(initial)**	t_2 **(prior-grounded)**	t_3 **(delta-focused)**
Talent	23/48/55	+8/+6/−8	+10/+3/−2
Market	25/65/78	+7/+10/−18	+3/+1/−1
Financial	10/85/95	**+25**/+3/−12	**+20**/+5/−3
IP	44/52/40	+6/+8/−5	+15/+4/−2
Regulatory	**56**/35/43	+3/+3/−3	+1/+1/−1
Total	**158/285/311**	**+49/+30/−46**	**+49/+14/−9**

complementary roles of temporal awareness: maintaining the validity of accumulated evidence and dynamically reallocating research effort as information needs evolve. Consequently, longitudinal research becomes a process of targeted evidence refinement rather than repeated exhaustive retrieval, achieving a 31% reduction in overall search effort while preserving attention to emerging signals.

5 Related Work

Agentic systems and deep research. Agentic system research spans from structured workflows for reasoning, optimization, and coding [63, 12, 43, 25, 35, 60, 46, 20, 47, 42, 9, 51, 36, 15, 58] to autonomous reasoning–action loops [52]. Multi-agent research advances from role-based collaboration and simulated interaction [21, 30, 5, 29] toward topology design, orchestration, memory, and fault tolerance [11, 31, 55, 14, 33, 57, 41]. Deep research extends these capabilities to sustained information seeking, as surveyed and benchmarked in [18, 54, 27, 3]. Representative systems decompose questions [7], optimize search through reinforcement learning [59, 19, 13, 39, 38], interpretable reasoning [48], or coordinate planning and evidence flows [23, 62, 28, 16]; web navigation and multimodal agents broaden the available evidence [44, 17]. These advances primarily concern how agents execute research. Our D/A distinction instead formalizes which evidence can be acquired deterministically and how it should guide subsequent exploration.

Epistemic reliability and domain-specific evaluation. Reliability research spans reflective and corrective retrieval, including multi-agent fact-checking [2, 50, 24, 61], alongside epistemic awareness, verification, uncertainty estimation, and context management [34, 26, 1, 49]. Such techniques strengthen evidential support, yet well-supported claims need not form a complete assessment: critical facts may remain unexplored, sources may conflict, and records may become stale. Domain agents address finance [53, 6, 10, 45, 4, 37], enterprise intelligence [56, 40], and law [8, 22], with deep research increasingly serving analytical workflows [32]. Enterprise evaluation particularly requires reconciling authoritative records with less structured signals while preserving traceability and acknowledging unknowns. Z-AXIS addresses this through D-conditioned coverage, evidence admission, and cross-dimensional reconciliation, coupling factual reliability with research completeness.

6 Conclusion

We introduce the *D/A Separation Principle*, which partitions the information space into programmable ($\mathcal{D}$) and emergent ($\mathcal{A}$) acquisition zones, and instantiated it as Z-AXIS, a three-phase agentic cascade for enterprise evaluation. By front-loading deterministic grounding as both epistemic anchor and investigative compass, Z-AXIS directs agentic exploration exclusively toward residual, high-value questions while enforcing admissibility filtering against the factual horizon. Experiments across five industries confirm that this architectural discipline yields superior factual accuracy, epistemic reliability, and coverage, with reliability governed by the architecture rather than backbone scale.

Broader impacts and limitations. Z-AXIS can support decision-making in investment due diligence, or credit assessment. Risks include over-reliance on automated evaluation and potential misuse for competitive intelligence. Current limitations include dependency on domain-specific structured data sources for $\mathcal{D}$ construction and evaluation confined to enterprise assessment, with varying availability of structured data across jurisdictions.

7 Acknowledgments

This work is supported by National Natural Science Fund of China (U25A20527), in part by the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX025G10).

References

[1] Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. In *Advances in Neural Information Processing Systems 37, NeurIPS 2024*, 2024. 10

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. 10

[3] Elad Ben Avraham, Changhao Li, Ron Dorfman, Roy Ganz, Oren Nuriel, Amir Dudai, Aviad Aberdam, Noah Flynn, Elman Mansimov, Adi Kalyanpur, and Ron Litman. DREAM: Deep research evaluation with agentic metrics. *arXiv preprint arXiv:2602.18940*, 2026. 10

[4] Tianshi Cai, Guanxu Li, Nijia Han, Ce Huang, Zimu Wang, Changyu Zeng, Yuqi Wang, Jingshi Zhou, Haiyang Zhang, Qi Chen, Yushan Pan, Shuihua Wang, and Wei Wang. FinDebate: Multi-agent collaborative intelligence for financial analysis. In *Proceedings of the 10th Workshop on Financial Technology and Natural Language Processing*, 2025. 10

[5] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. 10

[6] Yuemin Chen, Feifan Wu, Jingwei Wang, Hao Qian, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, and Meng Wang. Knowledge-augmented financial market analysis and report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, EMNLP 2024*, 2024. 10

[7] Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. MindSearch: Mimicking human minds elicits deep AI searcher. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, 2025. 1, 10

[8] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, et al. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7933–7962, 2024. 10

[9] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *Forty-first International Conference on Machine Learning, ICML 2024*, pages 13481–13544. PMLR, 2024. 10

[10] Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. FinEval: A chinese financial domain knowledge evaluation benchmark for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics, NAACL 2025*, 2025. 1, 10

[11] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*, 2023. 1, 10

[12] Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, Li Zhang, Lingyao Zhang, Min Yang, Mingchen Zhuge, Taicheng Guo, Tuo Zhou, Wei Tao, Robert Tang, Xiangtao Lu, Xiawu Zheng, Xinbing Liang, Yaying Fei, Yuheng Cheng, Yongxin Ni, Zhibin Gou, Zongze Xu, Yuyu Luo, and Chenglin Wu. Data interpreter: An LLM agent for data science. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025. 10

[13] Chen Hu, Haikuo Du, Heng Wang, Lin Lin, Mingrui Chen, Peng Liu, Ruihang Miao, Tianchi Yue, Wang You, Wei Ji, et al. Step-deepresearch technical report. *arXiv preprint arXiv:2512.20491*, 2025. 10

[14] Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, et al. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation. *arXiv preprint arXiv:2505.23885*, 2025. 1, 10

[15] Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025. 10

[16] Yusong Hu, Runmin Ma, Yue Fan, Jinxin Shi, Zongsheng Cao, Yuhao Zhou, Jiakang Yuan, Shuaiyu Zhang, Shiyang Feng, Xiangchao Yan, et al. Flowsearch: Advancing deep research with dynamic structured knowledge flow. *arXiv preprint arXiv:2510.08521*, 2025. 10

[17] Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Xu Tang, Yao Hu, Shaohui Lin, Philip Torr, Feng Zhao, and Wanli Ouyang. Vision-DeepResearch: Incentivizing DeepResearch capability in multimodal large language models. *arXiv preprint arXiv:2601.22060*, 2026. 10

[18] Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, Kun Shao, and Jun Wang. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025. 10

[19] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Proceedings of the Conference on Language Modeling, COLM 2025*, 2025. 10

[20] Boyan Li, Yuyu Luo, Chengliang Chai, Guoliang Li, and Nan Tang. The dawn of natural language to sql: Are we fully ready? *arXiv preprint arXiv:2406.01265*, 2024. 10

[21] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems 36, NeurIPS 2023*, 2023. 10

[22] Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344, 2025. 10

[23] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*, 2025. 1, 10

[24] Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See-Kiong Ng, and Tat-Seng Chua. FACT-AUDIT: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025*, 2025. 10

[25] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36, NeurIPS 2023*, 2023. 10

[26] Nahema Marchal, Stephanie Chan, Matija Franklin, Manon Revel, Geoff Keeling, Roberta Fischli, Bilva Chandra, and Iason Gabriel. Architecting trust in artificial epistemic agents. *arXiv preprint arXiv:2603.02960*, 2026. 10

[27] Saroj Mishra, Suman Niroula, Umesh Yadav, Dilip Thakur, Srijan Gyawali, and Shiva Gaire. SoK: Agentic retrieval-augmented generation (RAG): Taxonomy, architectures, evaluation, and research directions. *arXiv preprint arXiv:2603.07379*, 2026. 10

[28] Bo Pan, Lunke Pan, Yitao Zhou, Qi Jiang, Zhen Wen, Minfeng Zhu, and Wei Chen. InterDeepResearch: Enabling human-agent collaborative information seeking through interactive deep research. *arXiv preprint arXiv:2603.12608*, 2026. 10

[29] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023*, 2023. 10

[30] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, 2024. 10

[31] Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, 2025. 1, 10

[32] Matthew Russo and Tim Kraska. Deep research is the new analytics system: Towards building the runtime for AI-driven analytics. In *Conference on Innovative Data Systems Research, CIDR 2026*, 2026. 10

[33] Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or RL is suboptimal. In *Forty-second International Conference on Machine Learning, ICML 2025*, 2025. 10

[34] Jiaqi Shao, Yuxiang Lin, Munish Prasad Lohani, Yufeng Miao, and Bing Luo. Do LLM agents know how to ground, recover, and assess? A benchmark for epistemic competence in information-seeking agents. *arXiv preprint arXiv:2509.22391*, 2025. 10

[35] Zifan Song, Yudong Wang, Wenwei Zhang, Kuikun Liu, Chengqi Lyu, Demin Song, Qipeng Guo, Hang Yan, Dahua Lin, Kai Chen, et al. Alchemistcoder: Harmonizing and eliciting code capability by hindsight tuning on multi-source data. *Advances in Neural Information Processing Systems*, 37:2185–2214, 2024. 10

[36] Zifan Song, Xiao Gong, Guosheng Hu, Shuguang Dou, Qingsong Zhao, and Cairong Zhao. Dpl++: Advancing the network performance via image and label perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 10

[37] Zifan Song, Kaitao Song, Guosheng Hu, Ding Qi, Junyao Gao, Xiaohua Wang, Dongsheng Li, and Cai Zhao. Trade in minutes! rationality-driven agentic system for quantitative financial trading. In *International Conference on Learning Representations*, volume 2026, pages 27140–27156, 2026. 10

[38] Shiqian Su, Sen Xing, Xuan Dong, Muyan Zhong, Bin Wang, Xizhou Zhu, Yuntao Chen, Wenhai Wang, Yue Deng, Pengxiang Zhu, et al. Miroflow: Towards high-performance and robust open-source agent framework for general deep research tasks. *arXiv preprint arXiv:2602.22808*, 2026. 10

[39] Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025. 10

[40] Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. AutoML-Agent: A multi-agent LLM framework for full-pipeline AutoML. In *Forty-second International Conference on Machine Learning, ICML 2025*, 2025. 10

[41] Jen tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael R. Lyu, and Maarten Sap. On the resilience of LLM-based multi-agent collaboration with faulty agents. In *Forty-second International Conference on Machine Learning, ICML 2025*, 2025. 10

[42] Xiaohua Wang, Kaitao Song, Xu Tan, Huiqiang Jiang, Chengruidong Zhang, Yongliang Shen, Cen Lu, Zihao Li, Zifan Song, Caihua Shan, et al. Chain-of-model learning for language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 10

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35, NeurIPS 2022*, 2022. 1, 10

[44] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, Ding-Chu Zhang, Zekun Xi, Xiangru Tang, et al. Webdancer: Towards autonomous information seeking agency. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 10

[45] Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024. 10

[46] Yupeng Xie, Yuyu Luo, Guoliang Li, and Nan Tang. HAIChart: Human and AI paired visualization system. *Proc. VLDB Endow.*, 17(11):3178–3191, 2024. 10

[47] Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. DeepSeek-Prover: Advancing theorem proving in LLMs through large-scale synthetic data. In *Advances in Neural Information Processing Systems 37, NeurIPS 2024*, 2024. 10

[48] Jiaqi Xu, Cuiling Lan, Xuejin Chen, and Yan Lu. Stepwise think-critique: A unified framework for robust and interpretable llm reasoning. *arXiv preprint arXiv:2512.15662*, 2025. 10

[49] Renjun Xu and Yang Yan. Agent skills for large language models: Architecture, acquisition, security, and the path forward. *arXiv preprint arXiv:2602.12430*, 2026. 10

[50] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. 2024. 10

[51] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. 10

[52] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 10

[53] Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024. 1, 10

[54] Yipeng Yu. Deep research of deep research: From transformer to agent, from AI to AI for science. *arXiv preprint arXiv:2603.28361*, 2026. 10

[55] Sukwon Yun, Jie Peng, Pingzhi Li, Wendong Fan, Jie Chen, James Zou, Guohao Li, and Tianlong Chen. Graph-of-agents: A graph-based framework for multi-agent LLM collaboration. In *The Thirteenth International Conference on Learning Representations, ICLR 2026*, 2026. 10

[56] Yuhui Yun, Huilong Ye, Jingfeng Deng, Ruojia Li, Xinru Li, Li Li, and Haoyi Xiong. Eicopilot: Search and explore enterprise information over large-scale knowledge graphs with llm-driven agents. In *2025 IEEE International Conference on Big Data (BigData)*, pages 2689–2696. IEEE, 2025. 10

[57] Guibin Zhang, Muxin Fu, Kun Wang, Guancheng Wan, Miao Yu, and Shuicheng YAN. G-memory: Tracing hierarchical memory for multi-agent systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 10

[58] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025. 10

[59] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*, 2025. 1, 10

[60] Li Zhong, Zilong Wang, and Jingbo Shang. Debug like a human: A large language model debugger via verifying runtime execution step by step. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 851–870. Association for Computational Linguistics, 2024. 10

[61] Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. HaDeMiF: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, 2025. 10

[62] Bin Zhu, Qianghuai Jia, Tian Lan, Junyang Ren, Feng Gu, Feihu Jiang, Longyue Wang, Zhao Xu, and Weihua Luo. Marco DeepResearch: Unlocking efficient deep research agents via verification-centric design. *arXiv preprint arXiv:2603.28376*, 2026. 10

[63] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piekos, Aditya A. Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanic, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. Mindstorms in natural language-based societies of mind. *Computational Visual Media*, 11(1):29–81, 2025. 10

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We confirm that the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of the work in our conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided implementation details in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material.?

Answer: [Yes]

Justification: We have provided accessible qualitative results in our supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided implementation details in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have provided implementation details and failures in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided implementation details in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed both potential positive societal impacts and negative societal impacts of the work in our conclusion.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We use LLMs solely for checking grammar and polishing writing. Importantly, LLMs did not contribute to the conception of the research problem or the development of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.